Journal of Psychology and Behavioral Science
June 2014, Vol. 2, No. 2, pp. 205-220
ISSN: 2374-2380 (Print) 2374-2399 (Online)

Published by American Research Institute for Policy Development

Internal Consistency: Do We Really Know What It Is and How to Assess It?

Wei Tang[1], Ying Cui[2], Oksana Babenko[2]

Abstract

The term "internal consistency" of a test has been used widely but defined controversially in the field of psychometrics. In theoretical and practical research, internal consistency has had different meanings, such as homogeneity, average interitem correlation, general factor saturation, and internal consistency reliability. Lack of an explicit definition of internal consistency has posed difficulties in concept use and interpretation of results as well as hampered the development of new and better indices for measuring it. Building on the review of various meanings and measures of internal consistency, the present study attempts to provide an explicit definition of internal consistency, together with recommendations of appropriate measures for assessing it.

keywords: internal consistency, coefficient alpha, reliability, homogeneity

1. Introduction

The term "internal consistency" has been widely used but controversially defined. Cronbach (1951) used the terms "internal consistency" and "homogeneity" interchangeably stating that "an internally consistent or homogeneous test should be independent of test length" (p. 323). However, Revelle (1979) defined internal consistency as the extent to which all of the items of a test measure the same construct, that is, the general factor saturation.

[1] Department of Educational Psychology, University of Alberta, 6-102 Education North, University of Alberta, Edmonton, AB T6G 2G5, Canada. E-mail: wtang3@ualberta.ca
[2] Department of Educational Psychology, University of Alberta.

Yet other researchers (Green, Lissitz, & Mulaik, 1977; McDonald, 1981; Miller, 1995; Schmitt, 1996) used the term "internal consistency" to refer to the interrelatedness of items, and distinguished internal consistency from homogeneity by claiming that homogeneity refers to the unidimensionality of a set of test items. Another wide-spread use of internal consistency is related to the reliability estimate of a test based on a single administration, which is traditionally called internal consistency reliability (Haertel, 2006). In this sense, internal consistency has been often used to denote a group of methods that are intended to estimate reliability of a single-administrated test (Hattie, 1995).

As illustrated above, the term "internal consistency" is associated with different meanings. Just as Sijtsma (2009) concluded, "Internal consistency has not been defined that explicitly, far from it" (p. 114) and for a long time, this has posed difficulties in its use and interpretation of results. Further, the confusion in the concept of internal consistency may have also caused problems in understanding, measuring, and applying other related psychometric test properties such as internal consistency reliability, as well as hampered the development of new and better indices for measuring internal consistency. Therefore, an explicit and complete definition of the term "internal consistency" is highly needed. To better understand and define this term, various definitions or interpretations of the term "internal consistency" are reviewed and discussed in the following sections, together with the measures for assessing it. Following the review, an explicit definition of internal consistency is proposed. Numeric examples are also provided for the purpose of comparing the performances of different measures or indices of internal consistency and making recommendations on their use.

2. What Is Internal Consistency?

2.1. Internal Consistency = Average Interitem Correlation?

Conceptually, the internal consistency of a test indicates whether items on a test (or a subscale of a composite test), that are intended to measure the same construct, produce consistent scores. If, for example, ten items are designed to measure the same construct, an individual should answer these items in the same way, which would suggest that the test has internal consistency.

Thus, some researchers (Cortina, 1993; Cronbach, 1951) defined internal consistency as a measure based on the degree of bivariate correlations between different items on the same test (or the same subscale of a composite test). Since the correlations between items, most often than not, vary in size, using the average interitem correlation is a simple and direct approach to capture the degree of correlation between different items on a test. To the authors' knowledge, Cronbach (1951) is the first who proposed to use the average interitem correlation to measure internal consistency and developed a corresponding index: Cronbach's $\bar{r}_{ij}$. Cronbach's $\bar{r}_{ij}$ was derived by applying the Spearman-Brown formula (Brown, 1910; Spearman, 1910) to coefficient alpha, thereby estimating the mean of the correlations between items. Since then, internal consistency has been interpreted by some researchers and practitioners (e.g., Briggs & Cheek, 1986; Nunnally, 1978) as the average interitem correlation and assessed using Cronbach's $\bar{r}_{ij}$.

However, we should be aware that the concept of "consistent scores" does not necessarily mean the scores are identical or similar across items. For example, on a psychological test with two items, if all respondents agreed with the statement "I like to study with a partner" and disagreed with the statement "I hate to study with a partner", this would indicate the perfect internal consistency of the test. Although the two items are negatively correlated, the degree of the correlation is as high as 1. Suppose there are more items similar to the above two items added into the test, the average interitem correlations could be close to zero because positive interitem correlations can cancel out negative ones. To overcome this problem, practitioners either change the polarity of negative items into positive through recoding the observed scores or use the mean of absolute values of the correlations between items as a measure of internal consistency.

One disadvantage of the average interitem correlation is that this measure is influenced by the presence of extreme correlation values. In addition, it does not reflect the variability among the interitem correlations. Therefore, the use of the average interitem correlation is problematic when the correlations between items follow a skewed distribution, in particular when some extreme correlation values are observed. Revelle and Zinbarg (2009) also noted:

> The problem with interrelatedness is that it does not differentiate between the case in which each item is related to only a small proportion of the other items in the test from the case in which each item is related to every or nearly every other item in the test. (p. 152)

2.2. Internal Consistency = General Factor Saturation (GFS) or Closeness to Unidimensionality?

Revelle (1979) and Sijtsma (2009) offered another definition of internal consistency, which was an extent to which all of the items on a test measure the same construct, and coined it as the general factor saturation (Revelle, 1979). The advantage of this perspective over the notion of the average interitem correlation is that the general factor saturation is not affected by the skewness of the distribution of item correlations or extreme values of interitem correlations. Further, using the average interitem correlation to measure internal consistency is not appropriate for a multidimensional test that contains distinct subtests or subscales. Whereas the ideal measurement is for all items on a test to measure one latent trait or factor, this is hard to achieve in practice. For example, International English Language Testing System (IELTS, British Council, 1980) has four subscales which measure four distinctive skills: listening, reading, writing, and speaking. Although the test is designed to measure proficiency in English as a second language (higher order factor), each subscale measures a distinct trait or skill (lower level factor). This is not uncommon in psychological or educational testing, and hence the assumption of unidimensionality of a test is often violated. If subscales happen to be highly correlated, this suggests that the subscales, although originally intended to measure different traits, in fact, measure one trait. In this case, the test is factorially homogeneous (Revelle, 1979), with high correlations among the subscales indicating the presence of a higher order factor or only one general factor. Internal consistency is, thus, interpreted as the general factor saturation, which relates to homogeneity and explains why some researchers (Cronbach, 1951; Lord & Novick, 1968; Revelle, 1979) have used homogeneity and internal consistency interchangeably.

With respect to the measures of internal consistency when it is defined as general factor saturation, different approaches have been recommended. Revelle (1979) recommended using beta, which was defined as the lowest split-half reliability estimate, and pointed out that alpha tended to overestimate the general factor saturation while beta could give a more appropriate estimate of the general factor saturation in the case of in the case of "a lumpy test" (i.e., a test with several large group factors). To find the lowest split-half reliability estimate requires splitting the test in half in all possible ways, which is impractical when the test is, for example, made of 20 items. Because analytic methods will not work in that case, a heuristic procedure is needed to estimate beta. Several heuristic methods and software programs have been developed to address the estimation problem. For example, ICLUST (Revelle, 1977, 1979), a program for hierarchical cluster analysis, allows estimating beta. Recently, some researchers (Revelle & Zinbarg, 2009, Yang & Green, 2009, Zinbarg, Revelle, Yovel, & Li, 2005) recommended using ω_h (hierarchical coefficient omega, McDonald, 1999) as a measure of internal consistency. The ω_h indicates the extent to which all the items on a test measure the general factor and is defined as the ratio of the general factor variance to the total variance of the test (see Zinbarg et al., 2005 for the estimation of ω_h).

However, Sijtsma (2009) suggested that the internal consistency of a test should be assessed by determining the degree of closeness of the covariance/correlation matrix to unidimensionality, which was estimated using MRFA (minimum rank factor analysis, Ten Berge & Kiers, 1991). In MRFA, the closeness to unidimensionality is assessed using the ratio of the first eigenvalue to the sum of all the eigenvalues of the estimated true score covariance matrix (Ten Berge & Sočan, 2004). Expressed in percentages, Sijtsma (2009) referred to this ratio as the explained common variance (ECV). Nevertheless, Sijtsma (2009, p. 114) also noted that an internally consistent test is "psychologically interpretable" although this does not mean "that all items be factorially similar" (see Cronbach, 1951, p. 320).

2.3. Internal Consistency = Internal Consistency Reliability?

Internal consistency has also been used as a synonym of internal consistency reliability and thus some indices for measuring internal consistency reliability, in particular coefficient alpha (Cronbach, 1951), have been widely used.

However, some researchers (Cortina, 1993; Green et al., 1977; Nunnally, 1978) have pointed out the inappropriateness of using internal consistency reliability coefficients to measure internal consistency, stating that internal consistency is not an estimate of reliability. For example, Nunnally (1978) wrote:

> Estimates of reliability based on the average correlation among items within a test are said to concern the "internal consistency". This is partly a misnomer, because the size of the reliability coefficient is based on both the average correlation among items (the internal consistency) and the number of items. Coefficient alpha is the basic formula for determining the reliability based on internal consistency. (p. 229-230)

To better understand and distinguish between internal consistency and internal consistency reliability, we need to review the definition of reliability first. In the classical test theory, the term reliability was initially defined by Spearman (1904) as the ratio of true score variance to observed score variance. The estimation of reliability requires data from repeated testing, but such data are rarely available in practice. Thus, reliability is usually estimated based on scores from a single test administration, and is referred to as internal consistency reliability (Cronbach, 1951). Specifically, internal consistency reliability refers to the consistency of behavior within a very limited time interval, i.e., the time interval during which the items in the test are being responded to (Horst, 1953). From this point of view, the concept of internal consistency reliability seems to be similar to that of internal consistency. However, a test that is not internally consistent could be reliable. In other words, a high value of reliability does not guarantee a high level of internal consistency.

In contrast to reliability, internal consistency of a test is independent of its length (Cronbach, 1951). For example, a test may have unchanging internal consistency but increasing internal consistency reliability as the test length gradually increases from 10 items to, for instance, 20 items. Therefore, internal consistency reliability coefficients should not be used as indices of internal consistency because reliability estimates are functions of test length, that is, reliability increases as a test becomes longer. If test developers need to increase the reliability of a test they may just increase the number of items but this does not necessarily change the internal consistency of the test. In short, internal consistency is neither reliability nor internal consistency reliability.

2.4. An Explicit Definition of Internal Consistency

Based on the above review, internal consistency can, thus, be defined as a psychometric property of a test that is (a) associated with the degree of interitem correlations and the general factor saturation, and (b) independent of test length. It can be seen that internal consistency has a stronger connection with validity than reliability when the formulas of construct validity and reliability are compared (Judd, Smith, & Kidder, 1991; Streiner, 2003):

$$\text{Reliability}=\frac{\sigma_{CI}^2+\sigma_{SE}^2}{\sigma_T^2} \quad (1)$$

and

$$\text{Validity}=\frac{\sigma_{CI}^2}{\sigma_T^2}, \quad (2)$$

where σ_{CI}^2 is the variance of the construct of interest (i.e., the variance accounted by general factor), σ_{SE}^2 the systematic error variance, and σ_T^2 the total test variance. Therefore, if internal consistency is defined as the general factor saturation, measuring internal consistency is no different from measuring construct validity. However, as discussed earlier, internal consistency is a more complex concept and contains more information about the internal structure of a test than the general factor saturation.

The review of existing definitions of the term internal consistency has revealed that different researchers have attached different meanings to internal consistency and subsequently proposed a variety of indices for measuring internal consistency, including Cronbach's $\bar{r}_{ij}$ (Cronbach, 1951), Revelle's beta (Revelle, 1979), McDonald's ω_h (Zinbarg, Revelle, Yovel, & Li, 2005; Revell & Zinbarg, 2009), and Sijtsma's ECV (Sijtsma, 2009). When assessing the performance of these indices against the widely used coefficient alpha (Cronbach, 1951), the researchers demonstrated that these indices performed better than coefficient alpha.

However, none of the studies compared these indices simultaneously, and thus no explicit guideline has been developed yet. Therefore, the objective of the present study is to fill in this research gap by comparing these recommended indices for assessing internal consistency.

3. Method

To determine which of the recommended indices are appropriate measures of internal consistency, these indices were compared using hypothetical examples. The main advantage of using hypothetical data is that the internal structure of a test is known. Moreover, tests with various internal structures can be analyzed and compared simultaneously, allowing for the detection of the patterns or trends in the performance of the indices and the factors that may affect their performance.

In the present study, the simulation conditions were manipulated to resemble the structure of real psychological tests. Although psychological measures vary in terms of content such as intelligence tests, personality scales, and interest inventories, they have several common structural features that can be used in data simulation. First, psychological tests often measure several latent traits or attributes rather than a single trait as it is assumed in the classical test theory. For example, questions on the PDS (Personal Data Sheet, Woodworth, 1920) intend to measure excessive anxiety, depression, abnormal fears, and impulse problems among others. Second, the traits to be measured by a psychological test tend to be interrelated because they are components of a more general construct.

Considering these common structural features of psychological tests, the bifactor model (Chen, West, & Sousa, 2006; Holzinger & Swineford, 1937; Rindskopf & Rose, 1988) was used in the present study for simulating data that were structurally representative of existing psychological tests. The bifactor model includes the general factor and group factors and is given by:

$$X_i = \lambda_i F + \sum_{k=1}^{k} \alpha_{ik} G_k + E_i, \quad (3)$$

where X_i is the observed score for the i^{th} item, λ_i the factor loading of the general factor F on the i^{th} item, α_{ik} the factor loading of the K^{th} group factor G_k on the i^{th} item, and E_i the residual or randomness component of the i^{th} item.

The general method was to generate simulated population covariance matrices for observed component scores.

For the comparison purpose, the unifactor or unidimensional model was also included in the analyses, with α_{ik} set to zero. In total, five different internal structures were generated: (a) unifactor data with equal general factor loadings λ_i for all items, (b) unifactor data with unequal general factor loadings λ_i and odd-numbered items having larger loadings than even-numbered items, (c) bifactor data with larger general factor loadings λ_i than group factor loadings α_{ik}, (d) bifactor data with equal general factor loadings λ_i and group factor loadings α_{ik}, and (e) bifactor data with smaller general factor loadings λ_i than group factor loadings α_{ik}. In order to generate a simple bifactor structure, only two group factors (i.e., two subscales) with equal loadings α_{ik} were considered for the three bifactor internal structures. The specific values for factor loadings are shown in Tables 1 and 2 in the results section.

Next, each of the five structures was crossed with three levels of test length (10, 20, and 40 items) to examine the effect of test length on the performance of internal consistency indices. Finally, two levels of the average interitem correlation (low and medium, respectively) were considered for the unifactor and the two subscales of bifactor data sets. In total, thirty conditions were generated in the R environment (R Core Team, 2012) to compare and contrast the performance of internal consistency indices, including coefficient alpha, Cronbach's $\bar{r}_{ij}$, Revelle's beta, McDonald's ω_h, and Sijtsma's ECV. The standard deviation (SD) of the correlations between items was also examined under each condition.

4. Results

4.1. Unifactor Data with Equal Loadings

As shown in the upper panel of Table 1, Cronbach's $\bar{r}_{ij}$ (C's $\bar{r}_{ij}$), which by definition is the mean of interitem correlations, was equal to the square of the general factor loadings (λ_i) for the items for unifactor data with equal loadings. Their values were 0.30 and 0.60, respectively, regardless of the test length. Next, alpha, beta, and ω_h were equal at each level of test length.

These indices were positively affected by the increases in test length and general factor loadings. Specifically, as the test got longer (i.e., more items) and the size of general factor loadings was increased from $\lambda_i = \sqrt{0.30}$ to $\lambda_i = \sqrt{0.60}$, alpha, beta, and ω_h tended to increase. The ECV was equal to 1.0, indicating the perfect unidimensionality of the test. As expected, the SD of the correlations between items was zero, given the condition of equal item factor loadings.

Table 1: Indices for Measuring Internal Consistency for Unifactor Data

	Test Length					
	10	20	40	10	20	40
Equal Loadings		$\lambda_i = \sqrt{0.30}$			$\lambda_i = \sqrt{0.60}$	
C's $\bar{r}_{ij}$	0.30	0.30	0.30	0.60	0.60	0.60
alpha	0.81	0.90	0.95	0.94	0.97	0.98
beta	0.81	0.90	0.95	0.94	0.97	0.98
ω_h	0.81	0.90	0.95	0.94	0.97	0.98
ECV	1.00	1.00	1.00	1.00	1.00	1.00
SD	0.00	0.00	0.00	0.00	0.00	0.00
Unequal Loadings		$\lambda_i = \sqrt{0.62}$ and	$\lambda_j = \sqrt{0.10}$	$\lambda_i = \sqrt{0.92}$ and	$\lambda_j = \sqrt{0.35}$	
C's $\bar{r}_{ij}$	0.30	0.30	0.30	0.60	0.60	0.60
alpha	0.81	0.90	0.95	0.94	0.97	0.98
beta	0.68	0.74	0.78	0.89	0.92	0.93
ω_h	0.83	0.90	0.95	0.94	0.97	0.99
ECV	1.00	1.00	1.00	1.00	1.00	1.00
SD	0.18	0.19	0.19	0.19	0.20	0.20

Note. For equal loadings, λ_i is the loadings of the ith item (i=1, 2, 3 ... n). For unequal loadings, λ_i and λ_j are the loadings of the ith item (i=1, 3 ... n-1) and the jth item (j=2, 4 ... n), respectively. C's $\bar{r}_{ij}$ is Cronbach's $\bar{r}_{ij}$; ω_h is hierarchical coefficient omega; ECV is explained common variance; SD is the standard deviation of the correlations between items.

4.2. Unifactor Data with Unequal Loadings

As shown in the lower panel of Table 1, Cronbach's $\bar{r}_{ij}$, or the average interitem correlations, were again 0.30 and 0.60, respectively, regardless of the test length and the differences in interitem correlations as caused by unequal factor loadings. In contrast to the results for the unifactor data with equal loadings, alpha, beta and ω_h performed quite differently under the condition of unequal loadings. Among the three indices, beta had the lowest values while ω_h had the highest values at each level of the test length, with the difference being the largest at the lower level of the average interitem correlation (i.e., 0.30). Alpha became closer to beta as the average interitem correlation increased from 0.30 to 0.60 and the test length increased from 10 to 20 and then to 40 items. Overall, beta became considerably lower, alpha remained unchanged, and ω_h increased trivially under the condition of unequal loadings, when compared with the condition of equal loadings. The ECV was still equal to 1.0, indicating the perfect unidimensionality of these data. As expected, under the condition of unequal loadings, the SD was not equal to zero as it was under the condition of equal loadings. Still, the test length had little effect on the change in the SD. For example, as the item number grew up from 10 to 20, the SD only increased by 0.01.

4.3. Bifactor Data with High GFS

For the bifactor data with high GFS, that is, the data with the higher general factor loadings than the group factor loadings, Cronbach's $\bar{r}_{ij}$ became slightly lower (0.27 and 0.54, respectively; Table 2, upper panel), when compared with the corresponding values under the condition of unifactor model, although each of the two subscales of the bifactor data had the average interitem correlations of 0.30 and 0.60 as initially specified. Accordingly, alpha also became smaller since alpha was mainly affected by the average interitem correlation and test length. Beta and ω_h were equal, although lower than alpha across all the conditions of the bifactor data with high GFS. The discrepancies between them and alpha became slightly larger as test length increased from 10 to 40 items, with a small effect of the average interitem correlation. The ECV was 0.92 (close to 1), which indicated the bifactor data with high GFS was rather close to being unidimensional. The SD was not zero, but negligibly small and independent of test length. It increased as the average interitem correlations became larger.

Table 2: Indices for Measuring Internal Consistency for Bifactor Data

	Test Length					
	10	20	40	10	20	40
High GFS	$\lambda_i = \sqrt{0.25}$; $\alpha_{ik} = \sqrt{0.05}$			$\lambda_i = \sqrt{0.50}$; $\alpha_{ik} = \sqrt{0.10}$		
C's $\bar{r}_{ij}$	0.27	0.27	0.27	0.54	0.55	0.55
alpha	0.79	0.88	0.94	0.92	0.96	0.98
beta	0.72	0.81	0.85	0.85	0.88	0.89
ω_h	0.72	0.81	0.85	0.85	0.88	0.89
ECV	0.92	0.92	0.92	0.92	0.92	0.92
SD	0.02	0.02	0.02	0.05	0.05	0.05
Medium GFS	$\lambda_i = \sqrt{0.15}$; $\alpha_{j1} = \sqrt{0.15}$			$\lambda_i = \sqrt{0.30}$; $\alpha_{ik} = \sqrt{0.30}$		
C's $\bar{r}_{ij}$	0.22	0.22	0.22	0.43	0.44	0.45
alpha	0.73	0.85	0.92	0.88	0.94	0.97
beta	0.51	0.58	0.62	0.61	0.64	0.65
ω_h	0.51	0.58	0.62	0.61	0.64	0.65
ECV	0.75	0.75	0.75	0.75	0.75	0.75
SD	0.07	0.07	0.07	0.15	0.15	0.15
Low GFS	$\lambda_i = \sqrt{0.05}$; $\alpha_{ik} = \sqrt{0.25}$			$\lambda_i = \sqrt{0.10}$; $\alpha_{ik} = \sqrt{0.50}$		
C's $\bar{r}_{ij}$	0.16	0.17	0.17	0.32	0.34	0.34
alpha	0.66	0.80	0.89	0.83	0.91	0.95
beta	0.20	0.24	0.26	0.26	0.27	0.28
ω_h	0.20	0.24	0.26	0.26	0.27	0.28
ECV	0.58	0.58	0.58	0.58	0.58	0.58
SD	0.12	0.12	0.12	0.25	0.25	0.25

Note. λ_i is the general factor loadings of the ith item (i=1, 2, 3 ... n). α_{ik} is the group factor loadings of the ith item (i=1, 3,... n-1) when k is equal to 1 and the group factor loadings of the ith item (i=2, 4,... n) when k is equal to 2. GFS represents the general factor saturation. C's $\bar{r}_{ij}$ is Cronbach's $\bar{r}_{ij}$; ω_h is hierarchical coefficient omega; ECV is explained common variance; SD is the standard deviation of the correlations between items.

4.4. Bifactor Data with Medium GFS

For the bifactor data with medium GFS, the general factor and group factor loadings were assigned with equal values. The loadings were $\sqrt{0.15}$ and $\sqrt{0.30}$, corresponding to the two levels (low and medium) of the average interitem correlations. Still, each of the two subscales of the bifactor data had the average interitem correlations of 0.30 and 0.60. Compared with the bifactor data with high GFS, Cronbach's $\bar{r}_{ij}$ decreased to 0.22 and around 0.44, respectively (see Table 2, middle panel). Alpha also decreased accordingly. Beta and ω_h were equal, and became much lower than alpha across all the conditions of the bifactor data with medium GFS. The discrepancies between them and alpha became even larger as test length increased from 10 to 40 items and/or the average interitem correlation rose from low to medium. The absolute values of their discrepancies were larger than those under the condition of the bifactor data with high GFS. The ECV was 0.75, which indicated the bifactor data with medium GFS was a little away from being unidimensional. The SD increased slightly as the GFS changed from high to medium, and further increased as the average interitem correlation became larger.

4.5. Bifactor Data with Low GFS

To generate the bifactor data with low GFS, the general factor and group factor loadings in the bifactor data with high GFS were switched. Cronbach's $\bar{r}_{ij}$ decreased even further to around 0.17 and 0.34, respectively (see Table 2, bottom panel). Alpha also continued to decrease. However, the decrease was trivial when the test length was large or when the average interitem correlation was at the medium level. Beta and ω_h values were equal but decreased remarkably. As shown in Table 2 (bottom panel), when the general factor loadings were $\sqrt{0.05}$, beta and ω_h values decreased by more than 70% of the values under the condition with high GFS and became much lower than alpha across all the conditions of the bifactor data with low GFS. Furthermore, beta and ω_h increased much less than alpha as the test length increased from 10 to 40 items and/or the level of the average interitem correlation increased from low to medium. The ECV was 0.58, which indicated the dimensionality of the bifactor data with low GFS was farther away from being unidimensional. The SD continued to increase as the GFS changed from medium to low, and the increase was boosted as the average interitem correlation grew up.

5. Discussion

For several decades, the term "internal consistency" of a test has been associated with different meanings, including homogeneity, interrelatedness, general factor saturation and internal consistency reliability. It poses difficulties in concept use and interpretation of results as well as hampering the development of new and better indices for measuring internal consistency. Based on the review of existing definitions, internal consistency can, thus, be defined as a psychometric property of a test that is (a) associated with the degree of interitem correlations and the general factor saturation, and (b) independent of test length.

In order to determine and evaluate which indices are appropriate for measuring internal consistency, the following three criteria were considered. The first criterion is the ability to reflect the degree of the general factor saturation of a test. Out of the six indices considered in the present study, only beta and ω_h, were shown to be able to clearly indicate the change in the degree of the general factor saturation in our numerical examples. The second criterion is the ability to reflect the degree of interrelatedness, that is, the degree of interitem correlations. Except for the ECV and SD, the considered indices were shown to be able to reflect the change in the degree of interitem correlations. The third criterion is the independence of test length. Alpha, beta, and ω_h were shown to depend on the test length manipulated in the present study (i.e., 10, 20, and 40 items). These indices tended to increase in value with an increase in the test length. Out of these three indices, alpha was found to be the most influenced by the test length. Although Cronbach's $\bar{r}_{ij}$, SD, and ECV were determined to be independent of the test length, these indices, if used alone, do not depict the whole picture of test internal consistency.

The results of the comparison of the six indices that are currently in use in the field of measurement and testing indicate that using a single measure to assess internal consistency is not sufficient, and thus, a combination of measures is recommended. In order to assess internal consistency, it is recommended to use ECV first to assess whether or not a test is unidimensional or close to unidimensionality, because the performances of other indices are affected by the closeness to unidimensionality. Although ECV performs well as a measure of the closeness to unidimensionality, it fails to reveal the change in interitem correlations.

Given this, it is recommended that when the ECV value is high (e.g., greater than 0.75 in this study), both the average degree of the correlations between items (Cronbach's $\bar{r}_{ij}$) and the SD of these correlations should be reported. These two indices function well in revealing the change in interitem correlations under the condition of data being close to unidimensionality. If ECV is medium or low (e.g., 0.75 or below in this study), beta and ω_h are recommended because these indices reflect the degree of the general factor saturation and overcome the shortcoming of average interitem correlation under the condition of data being heterogeneous or multidimensional.

In conclusion, the existence of multiple and controversial definitions of internal consistency, along with various indices to measure it, has presented difficulties in concept use and interpretation of results. The fact that none of the examined six indices is able to provide a complete picture of internal consistency when used alone, calls for the development of a new and better measure of internal consistency. Future research should address this long-standing need in the field of measurement and assessment.

6. References

Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. Journal of Personality, 54, 107-148.

British Council, (1980), International English Language Testing System (IELTS). Retrieved from http://www.ielts.org/default.aspx

Brown, W. (1910). Some experimental results in the correlation of mental abilities. British Journal of Psychology, 3, 296–322.

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality-of-life. Multivariate Behavioral Research, 41, 189-25.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. Journal of Applied Psychology, 78, 98-104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297-334.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. Educational and Psychological Measurement, 37, 827-838.

Haertel, E. H. (2006). Reliability. In Brennan, R. L. (Ed.). Educational Measurement. (4th ed.). National Council on Measurement in Education. American Council on Education. Westport, CT : Praeger Publishers.

Hattie, J. A. (1995). Methodology review: assessing unidimensionality. Applied Psychological Measurement, 9, 139-164.

Holzinger, K. J., & Swineford, F. (1937). The Bi-factor method. Psychometrika, 2, 41–54.

Horst, P. (1953). Correcting the Kuder-Richardson reliability for dispersion of item difficulties. Psychological Bulletin, 50, 371-374.

Judd, C. M., Smith, E. R., & Kidder, L. H. (1991). Research methods in social relations (6th ed.). New York: Harcourt Brace Jovanovich.

Lord, F. M., & Novick, M.R. (1968). Statistical theories of mental test scores. Menlo Park, CA: Addison-Wesley.

McDonald, R. P. (1981). The dimensionality of tests and items. British Journal of Mathematical and Statistical Psychology, 34, 100-117.

McDonald, R. P. (1999). Test theory: A unified treatment. Hillsdale: Erlbaum.

Miller, M. B. (1995). Coefficient alpha: A basic introduction form the perspectives of classical test theory and structural equation modeling. Structural Equation Modeling, 2, 255-273.

Nunnally, J. C. (1978). Psychometric theory (2nd ed.). New York: McGraw-Hill.

Revelle, W. (1977). ICLUST: A program for hierarchical cluster analysis. Northwestern University. Computing Center Document 482.

Revelle, W. (1979). Hierarchical cluster-analysis and the internal structure of tests. Multivariate Behavioral Research, 14 (1), 57-74.

Revelle, W., & Zinbarg, R. E. (2009). Coefficient alpha, beta, and the glb: Comments on Sijtsma. Psychometrika, 74, 145-154.

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. Multivariate Behavioral Research. 23, 51–67.

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Schmitt, N. (1996). Uses and abuses of coefficient alpha. Psychological Assessment, 8, 350-353.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika, 74, 107-120.

Spearman, C. (1904). The proof and measurement of association between two things. American Journal of Psychology, 15, 72-101.

Spearman, C. (1910). Correlation calculated from faulty data. British Journal of Psychology, 3, 271–295.

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. Journal of Personality Assessment. 80, 99-103.

Ten Berge, J. M. F., & Kiers, H. A. L. (1991). A numerical approach to the exact and the approximate minimum rank of a covariance matrix. Psychometrika, 56, 309-315.

Ten Berge, J. M. F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. Psychometrika, 69, 613-625.

Woodworth, R. S. (1920). Personal data sheet. Chicago: Stoelting.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α , Revelle's β, and McDonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. Psychometrika, 70, 123-133.